DOCUMENT RESUME

ED 258 988 TM 850 356

AUTHOR Reid, Jerry B.
TITLE Establishing Upper Limits for Item Ratings for the Angoff Method: Are Resulting Standards More 'Realistic'?
PUB DATE Apr 85
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Certification; Cutting Scores; *Difficulty Level; Interrater Reliability; *Item Analysis; Multiple Choice Tests; Standards; *Test Items
IDENTIFIERS *Angoff Methods; *Standard Setting

ABSTRACT

This report investigates an area of uncertainty in using the Angoff method for setting standards, namely whether or not a judge's conceptualizations of borderline group performance are realistic. Ratings are usually made with reference to the performance of this hypothetical group, therefore the Angoff method's success is dependent on this point. A technique whereby the conceptualization of borderline performance can be defined in reference to the performance of the total group is demonstrated. If total group performance is more familiar to judges, either through past experience or training provided during the standard setting activity, use of total group performance to provide a frame of reference within which to define group performance may lead to more realistic item ratings. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED258988

ESTABLISHING UPPER LIMITS FOR ITEM RATINGS FOR THE ANGOFF METHOD: ARE RESULTING STANDARDS MORE 'REALISTIC'?

Jerry B. Reid, Ph.D.
The American Registry of Radiologic Technologists

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Reid, Jerry B.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, April 1985.

TM 850356

ABSTRACT

AN AREA OF PARTICULAR UNCERTAINTY IN USING THE ANGOFF METHOD FOR SETTING STANDARDS IS WHETHER A JUDGE'S CONCEPTUALIZATIONS OF BORDERLINE GROUP PERFORMANCE ARE REALISTIC. SINCE ALL RATINGS ARE MADE WITH REFERENCE TO THE PERFORMANCE OF THIS HYPOTHETICAL GROUP, THE SUCCESS OF THE METHOD HINGES UPON THIS POINT. A TECHNIQUE IS DEMONSTRATED WHEREBY THE CONCEPTUALIZATION OF BORDERLINE PERFORMANCE CAN BE DEFINED IN REFERENCE TO THE PERFORMANCE OF THE TOTAL GROUP. IF TOTAL GROUP PERFORMANCE IS MORE FAMILIAR TO JUDGES, EITHER THROUGH PAST EXPERIENCE OR TRAINING PROVIDED DURING THE STANDARD SETTING ACTIVITY, USE OF TOTAL GROUP PERFORMANCE TO PROVIDE A FRAME OF REFERENCE WITHIN WHICH TO DEFINE BORDERLINE GROUP PERFORMANCE MAY LEAD TO MORE REALISTIC ITEM RATINGS.

Under the Angoff Method of standard setting (Angoff, 1978), judges are asked to estimate item difficulties with reference to a hypothetical group of borderline examinees. The success of the method hinges upon how "realistic" the individual judge's conceptualization of borderline performance is. Since no criterion measure exists for definitively describing the borderline group when using the Angoff Method (or else it would be used instead in setting the standard), what is realistic can only be defined very generally.

For the typical certification examination situation in which training programs are well established, the match between students selected into the programs and requirements of the profession is appropriate, and the match between the examination content and requirements of the profession is appropriate, it is expected that most examinees will perform well leading to a negatively skewed distribution of test scores. Based upon the perceptions of those most familiar with the requirements of the profession and with typical program graduates, it is realistic to expect that most examinees should be classified as "passers." Even with the above hurdles in place, it is also realistic to expect that some small percentage of program graduates will typically not possess sufficient mastery of the knowledge and skills to be classified as passers. A judge's ability to conceive of a borderline group at a reasonable point in the tail of this performance distribution and to rate individual items relative to this group determines the reasonableness of the standard derived. If reference points can be established to help guide a judge's conception of borderline performance, then the validity of the ratings may be enhanced.

If test items are appropriate for the total group of program graduates, it may be expected that most items will be relatively easy for this group (i.e., item difficulty indices will be high). If a subgroup could be isolated from this total group such that it was composed of that part of the score distribution located at the borderline of the mastery/non-mastery point, it would be realistic to expect that the items would be more difficult generally for this group and that item difficulty indices would be lower in most cases than for the total group. In this situation, the total group item difficulty indices may be thought of as an upper limit on the Angoff ratings for items.

Since total group performance is something with which judges may already be familiar or with which they can be made familiar (provided that the items to be rated or similar items have been previously administered), the estimation of total group item difficulty indices by judges prior to making Angoff ratings may act to provide a more familiar frame of reference within which to rate items for the hypothetical group of borderline examinees. Note that since judges are estimating total group item difficulty indices and are not being given actual observed item difficulties, judgements are not being contaminated by observed data (if such contamination is felt to be undesirable). The total group item difficulty ratings can be used also as a check on the ability of judges to estimate a known value in order to identify those judges who consistently make unrealistic ratings in order to eliminate or re-train them.

The objective of this study was to determine whether the estimation of total group item difficulty indices by judges prior to using the Angoff Method influences the resulting standard.

INSTRUMENTATION

Twenty multiple-choice items from a previously administered 250-item certification examination were selected for use in this study. The examination is administered to voluntary candidates for national certification after having completed an accredited two-year educational program. Items were selected to represent a range of content and item difficulty levels. A distribution of the observed item difficulty indices (expressed as the percentage of examinees answering the item correctly) for the twenty items based upon the responses of 3221 examinees is shown in Figure 1. A score distribution based upon responses of the 3221 examinees to this subset of twenty items is shown in Figure 2.

Items were reproduced on 35mm slides for presentation to the judges. Items were formatted with one item per slide with the correct answer indicated.

Two alternate sets of materials were developed describing the procedures to be followed in rating items. Both sets of materials contained the same introductory page describing the need for collecting data for use in establishing a defensible standard for certification examinations, the general procedures to be followed in the exercise (e.g., projection of the items, et cetera), the existence of two different sets of instructions, and the consequent need to make judgements independently, without consultation with neighboring judges. No mention of the concep' of borderline

examinees or other method-specific information was discussed on this sheet. Following a page requesting various demographic information and the page on which ratings were to be recorded, was a set of group-specific instructions.

The first set of materials (Group A Instructions) requested that the judge "estimate the percentage of graduates of approved educational programs taking the examination for the first time who would answer a question correctly on the first round of rating." Judges were further requested to make ratings in multiples of five (i.e., 100, 95, 90, 85, ...). The directions requested judges to make their ratings without reading further in their packets. A subsequent page in the Group A Instructions described the procedures for using the Angoff Method on the second round of ratings, again making ratings in multiples of five.

The alternate set of materials (Group B Instructions) contained the Angoff Method instructions to be followed on the first round of rating in exactly the same language as they appeared in the Group A Instructions. The set of procedures to be followed in the second round for the Group B Instructions requested that the Angoff Method be repeated. Figure 3 illustrates the difference between the two sets of instructions.

JUDGES

The group of judges was composed primarily of educators from the discipline represented by the test items. A small number of non-educators (administators and practitioners) from the discipline were included as well. The group was not necessarily representative of a panel that would be selected normally for an actual standard

setting procedure. The exercise was carried out primarily to inform the group as to general methods of standard setting used on the certification examination and to gather information on the feasibility of using the proposed technique operationally.

METHOD

The two sets of materials describing the rating processes were arranged in alternating fashion and distributed to the group, thereby dividing the group into two approximately equal-sized subgroups. The introductory page was read and discussed with the group as a whole. Judges were next requested to complete the demographic questions and to read the first set of instructions to themselves. Questions were responded to individually rather than being discussed with the entire group. Since the directions given to the two groups were different and the groups were not separated, any group discussion of the meaning of borderline or individual item ratings as would normally be carried out was precluded.

After all questions regarding instructions had been addressed, the items reproduced on slides were projected. As each item was projected, judges marked their ratings on a response sheet according to their set of directions.

Each group rated the set of items twice. The first time through the slides, Group B rated the items according to the conventional Angoff Method. Group A estimated the item difficulties based upon a hypothetical group composed of typical graduates of educational programs in the discipline taking the examination for the first time.

On the second presentation of the slides, Group B was asked to repeat the Angoff Method, reconsidering each item in light of their newly acquired experience with the method. Group A was asked to rate each item according to the conventional Angoff Method. The ratings made by a judge on the first round was available to that judge on the second round of ratings.

RESULTS

A summary of the results is shown in Table 1:

Table 1

	GROUP A		GROUP B	
	ROUND #1	ROUND #2	ROUND #1	ROUND #2
MEAN	82.91	59.64	72.91	70.76
S.D.	6.65	18.86	12.79	12.80
N	26	26	29	29

The primary comparison of interest was between the Round #2 ratings given by Group A and Group B. The comparison of the standards set by Group A with that for Group B on the second round indicated a statistically significant difference (p less than 0.05 based upon Behrens-Fisher technique necessitated by unequal sample sizes and unequal variances).

In addition to being statistically significant, the effect of the standards set by the two groups was significant from a practical standpoint when considering the percentage fail rate that would have

resulted if the twenty-item test had been used with each of the two standards. Based upon the distribution of scores (Figure 2) for this subset of twenty items the standard set by Group B (i.e., 15 items correct) would have resulted in a fail rate of 23% while the standard set by Group A (i.e, 12 items correct) would have resulted in a fail rate of 6%. However, part of this extreme difference in fail rates may be due to the coarseness of the distribution provided by only twenty items.

The mean estimated item difficulty index for the hypothetical total group of first-time examinees of 82.91 as made by Group A on Round #1 agrees favorably with the mean observed item difficulty value of 81.65 (based upon 3221 examinees). The correlation of the 20 mean estimated item difficulties from Group A on Round #1 with the 20 observed item difficulties was 0.73. Although the results of these group-based checks on the judges' ability to make realistic estimates of item difficulty indices for the total group of typical examinees are reasonable, the ability of individual judges to make realistic ratings varied considerably. The correlations for individual judges of estimated to observed indices varied from 0.01 to 0.65. The mean of the estimated item difficulty indices for individual judges ranged from 74 to 93.

While it was suspected that some judges in Group A might set Angoff ratings by simply subtracting a constant from their estimates of total group item difficulties, the data did not indicate that this occurred. The standard deviations of the arithmetic difference between the Round #1 ratings and Round #2 ratings for judges ranged from 3.80 to 18.91 with a mean standard deviation of 13.21.

DISCUSSION

Assuming that the results are replicable under conditions more typical of standard setting procedures in terms of representativeness of judges and discussion of ratings, the results do demonstrate that Angoff ratings are influenced by having judges estimate total group item difficulty indices prior to using the Angoff Method. However, such a demonstration does not prove that the resulting standards are more "realistic." In fact, the method may simply act to artificially depress ratings given with the Angoff Method. Since what is a realistic standard must always be a subjective decision in the absence of a criterion measure to define borderline performance, the judgement as to whether the technique leads to more realistic results must also be subjective. The reasoning that any information provided to judges which makes the conceptualization of the expected performance of the borderline group less ephemeral should lead to more appropriate judgments, argues that the resulting standard should be more realistic. Constraining borderline performance within an upper bound (total group performance) provides additional structure to a judge's task by narrowing the range of ratings available.

The fail rate for the certification examination from which the twenty items used in this study were taken has ranged between 11% and 14% over the last five years. Given that these fail rates have been considered "realistic" by this examination program, the standards established by both Group A and Group B are somewhat deviant although that of Group A is less so.

Having judges first estimate known indices (i.e., observed item difficulty values for a group of examinees representing the typical performance distribution) opens the possibility of identifying judges who seem unable to make realistic judgements. A judge incapable of making realistic judgements for a known (or knowable) group will presumably be equally unprepared to judge the hypothetical borderline group. Such identification would make either remediation or elimination of the judge from the study possible.

In a series of studies Lorge, et. al. (Lorge and Kruglov, 1952; Lorge and Kruglov, 1953; Lorge and Diamond, 1954; Lorge and Diamond 1954) investigated the ability of judges to estimate the difficulty of items. Their findings indicate that judges can generally estimate the relative difficulty of items quite well and that familiarization of judges with the observed difficulty indices of a representative set of items (i.e., training) can improve the validity of judgements regarding absolute item difficulty. In particular, such training can bring the performance of the most inaccurate judges up to that of the most accurate judges. By first orienting judges using this training technique, the number of judges that it would be necessary to eliminate from the standard setting activities should be minimized and the "upper limits" for Angoff ratings be made more realistic. Since the group of judges used in this study was composed primarily of educators, it was felt that training (i.e, familiarization of representative items) was not necessary. However, in retrospect, such training would probably have improved the estimates of total group item difficulty indices and led to more realistic upper bounds for the Angoff ratings.

CONCLUSIONS

While it does appear that having judges estimate item difficulty values based upon a group of typical examinees representative of the entire performance distribution prior to estimating item difficulty values for the hypothetical borderline group (Angoff Method) does affect the resulting standard set, whether the result is more "realistic" is a subjective issue. The rationale that providing judges with a more familiar framework within which to base the judgements of borderline performance argues that the standard will be more firmly anchored in reality.

It has been proposed that a "reality check" be included as part of the standard setting procedure (Livingston and Zieky, 1982). The proposed technique provides a potentially sensitive reality check in that the focus is on individual items as opposed to the final fail rates that would be produced by the results. Identifying judges who are initially unable to make realistic judgements regarding total group performance makes remediation efforts possible. This sensitization to total group performance should in turn make judgements regarding borderline performance more realistic.

REFERENCES

Angoff, W. H. Scales, Norms and Equivalent Scores. In R. L. Thorndike (ed.), Educational Measurement Washington, D.C., American Council on Education, 1971, pp. 514-515.

Livingston, S. A., Zieky, M. J. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Princeton, NJ: Educational Testing Service, 1982.

Lorge, I., Kruglov, L. A suggested technique for the improvement of difficulty prediction of test items. Educational and Psychological Measurement, 12, pp. 554-561.

Lorge, I., Kruglov, L. The improvement of estimates of test difficulty. Educational and Psychological Measurement, 13, pp. 34-46.

Lorge, I., Kruglov, L. The value of information to good and poor judges of item difficulty. Educational and Psychological Measurement, 14, pp. 29-33.

FIGURE 1: DISTRIBUTION OF OBSERVED ITEM DIFFICULTY VALUES (P+)

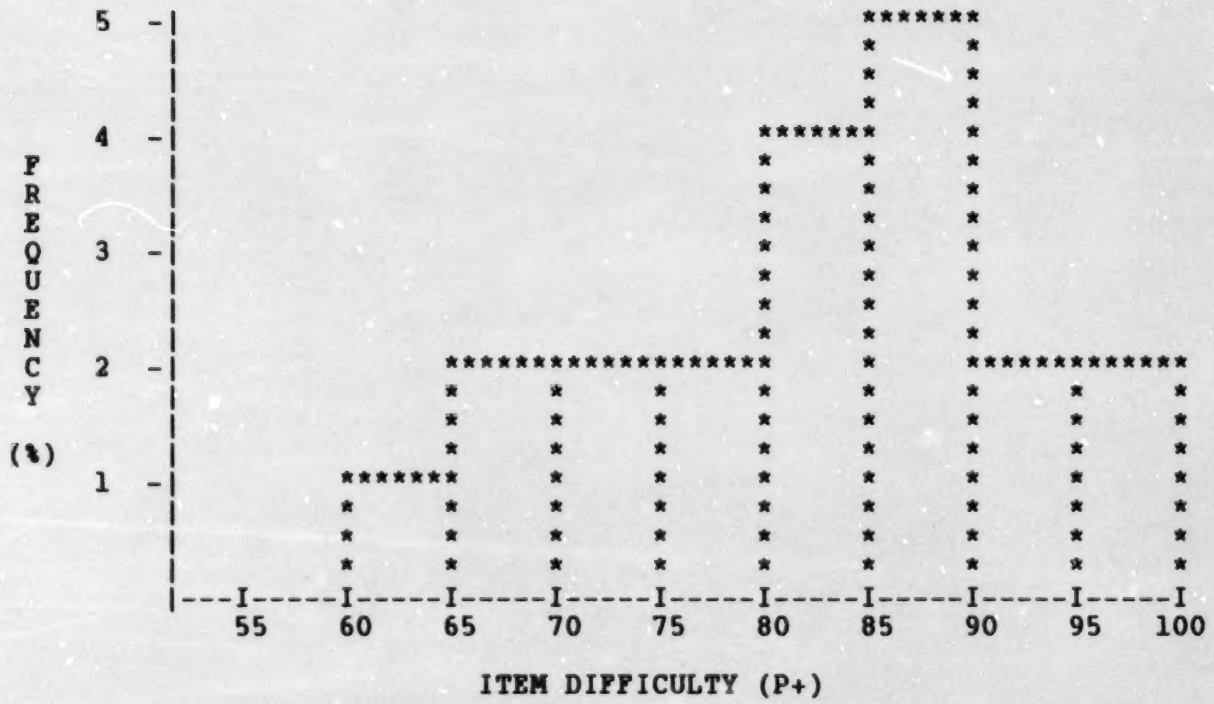

FIGURE 2: SCORE DISTRIBUTION FOR 20-ITEM TEST (N=3221)

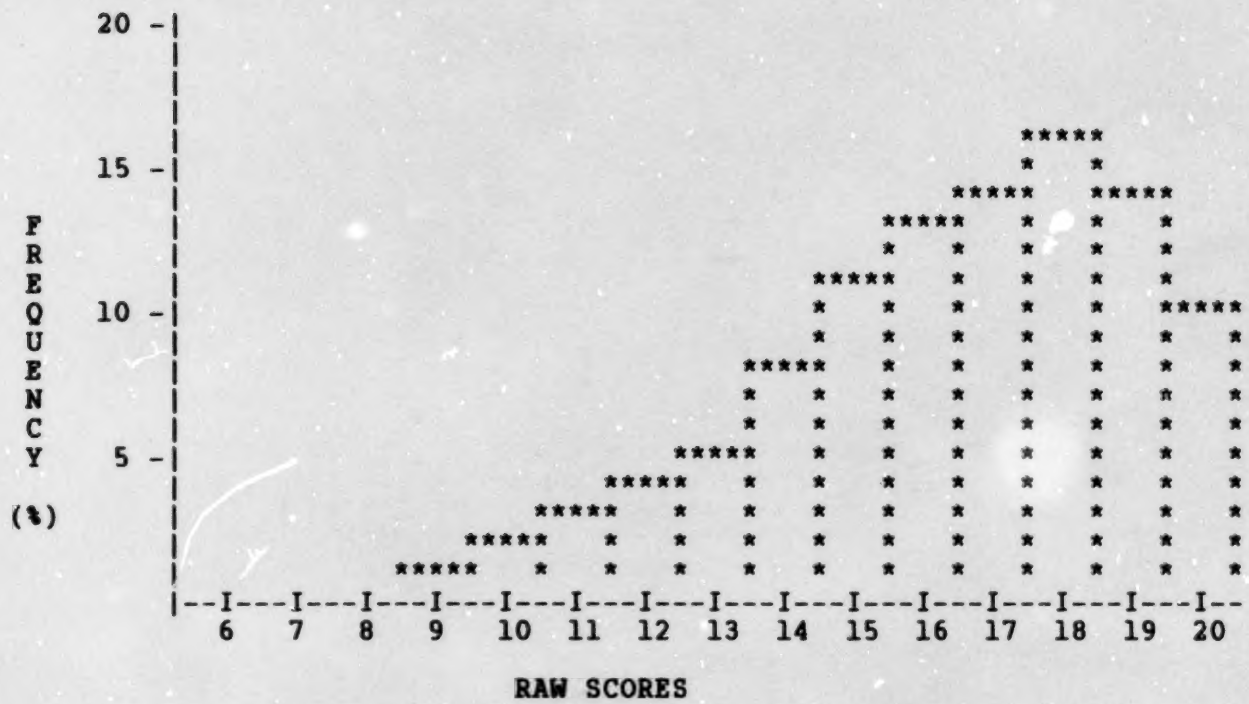

FIGURE 3: INSTRUCTION MATERIALS FOR JUDGES

PAGE #	GROUP A	GROUP B	RELATIONSHIP OF A & B
1	Introduction	Introduction	Same
2	Demographics	Demographics	Same
3	Rating Sheet	Rating Sheet	Same
4	Total Group Estimation Directions	Angoff Method Directions	Different
5	Angoff Method Directions	Angoff Method Directions	Same